Big Data Analytics Final Project

Movie Dataset

Professor Girish Srinivasan
STAT 580: AIT
George Mason University

Report by
Srashti Agrawal
G01079416
Date- 09/12/2018

Introduction:

Movies are becoming an important part of life. In the age of digital world, studying its role and impact on society is a challenging task. It entertains people and influence them and at the same time it also helps to make profit for film companies which motivates them to create more movies for people. So, I picked the dataset of movies to perform some general data analysis. My data doesn't include any database so didn't use SQL for data exploration instead used R for data scraping or web scraping.

The dataset is in Comma Separated Value format which contains 1773 entries and has 12 variables initially. After hot encoding it becomes 110 variables. It contains both numeric and character type of data.

Dataset contains different types of data types, some of the variable names are:

Quantitative: Rank(Continuous), Total Movie Earning, Theater, Years etc.
Qualitative: Movie_Name, Studio_Name, Genre(Nominal) , Director etc.

Data Collection:

The data is collected from different websites by scraping various movie websites such as IMDB Omdbapi, Rotten Tomato, Box Office Mojo & National Association of Theatre Owners.

1. *IMDB*[1]: IMDb (Internet Movie Database) is an online database of information related to films, television programs, home videos and video games, and internet streams.

2. *Rotten Tomato*[2]: American review-aggregation website for film and television.

3. *Box Office Mojo*[3]: Box Office Mojo is a website that tracks box office revenue in a systematic, algorithmic way.

4. *National Associated of Theatre Owners*[4]: The National Association of Theatre Owners (NATO) is a United States based trade organization whose members are the owners of movie theaters.

Requirements, resources needed:

Software requirements: -
R is used for analysis of dataset, data exploration, transforms, and Web scraping.

Hardware requirements:
All analysis for this project was performed with Apple laptop of the following characteristics: -

Data Ingestion:

I have used the '*rvest*' library (https://github.com/hadley/rvest) provided in R to scraped data from BoxOfficeMojo (https://www.boxofficemojo.com/yearly/). Rvest is inspired from BeautifulSoup package in R and uses *magrittr* to perform common web page scraping tasks.

I have also used the 'omdbapi' package, it is an independent package developed to utilize the omdb api (http://www.omdbapi.com/) to access IMdB movie data. The api provides extensive data on each of the movie such as genre, rating, director, actor names.

Data Description:

I choose top 100 earning movies for each year, as sufficient data is present for each of them and focus on blockbuster movies. I included only Hollywood movies that were in English as we are interested in domestic US Box Office results and thus were only looking at U.S. gross revenue, we choose movies from the year 2001 to 2018. The reason for placing the time period restriction was that we only wanted to include recent movies as it would have been difficult to compare movies from different eras. Over time, movie tastes would have changed, meaning the characteristics of a profitable movie would have also changed. Our model would be unable to take these changes into account.

In recent years, movies have generally become divided into two categories: blockbusters and independent movies. Studios have focused on relying on only a handful of extremely expensive movies every year to make sure they remain profitable. It is estimated that 80% of the industry's profits over the last decade is generated from just 6% of the films released; 78% of movies have lost money of the same time period.

I have also scraped yearly data such as Total earning by movies each year, average ticket price, total movies released in a year.

Year	Rank	Movie.Name	Studio.Name	Total.Movie.Earning.million	Theater	Ea
2018	1	Black Panther	BV	700059566	4084	
2018	2	Avengers: Infinity War	BV	678815482	4474	
2018	3	Incredibles 2	BV	608563044	4410	
2018	4	Jurassic World: Fallen Kingdom	Uni.	416769345	4485	
2018	5	Deadpool 2	Fox	318491426	4349	
2018	6	Mission: Impossible - Fallout	Par.	220159104	4395	
2018	7	Ant-Man and the Wasp	BV	216648740	4206	
2018	8	Solo: A Star Wars Story	BV	213767512	4381	
2018	9	Venom	Sony	212411166	4250	
2018	10	Dr. Seuss' The Grinch	Uni.	211619175	4141	
2018	11	A Star is Born	WB	195278597	3904	
2018	12	A Quiet Place	Par.	188024361	3808	
2018	13	Crazy Rich Asians	WB	173962956	3865	
2018	14	Bohemian Rhapsody	Fox	169219170	4000	
2018	15	Hotel Transylvania 3: Summer Vacation	Sony	167510016	4267	
2018	16	Halloween	Uni.	159211275	3990	
2018	17	The Meg	WB	143005856	4118	
2018	18	Fantastic Beasts: The Crimes of Grindelwald	WB	140187524	4163	

MetaData:

1. Year :: The year the movie is released.
2. Rank :: The rank of the movie in terms of earning for the corresponding year the movie is released.
3. Movie.Name :: The name of the movie.
4. Studio.Name :: The name of the studio producing the movie.
5. Total.Movie.Earning.million :: Total movie domestic (US) earning (in $)
6. Theater : Number of theatres the movie is released in.
7. Earning.Opening.Day.million :: Earning of the movie on the opening day.
8. Rated :: The rating provided to the movie for example, R, PG-13 etc.
9. Runtime.min :: The total length of the movie in mins.
10. Director :: The name of the director of the movie.

11. ImdB.names.used :: The name of the movie in IMDB website.
12. Genre :: The genre of the movie for example, Action, Fantasy etc
13. Total.Movies.Released :: Total number of movies released in the corresponding year.
14. Gross.Earning.million :: Total gross earning made by all the movies in the calendar year.
15. Total.Tickets.Sold.million :: Total tickets sold in the calendar year.
16. Avg.Ticket.Price :: Average ticket price in the calendar year.
17. Inflation.Adjustment :: Inflation adjustment factor with 2001 as the base year.
18. Adj.Total.Movie.Earning.million :: Inflation adjusted movie earning.
19. Adj.Total.Gross.Earning.million :: Inflation adjusted total gross earning in the calendar year.
20. Percentage.Earning :: Percentage of the total gross earning a movie made.
21. Adj.Earning.Opening.Day.million :: Inflation adjusted Opening data earning by a movie.

Data Issues:

As mentioned there are only 100 top grossing movies taken into account while more than 700 movies were released in hollywood alone released in theatre. I am considering only US box office earnings as the earning are made available by studios and provided to general people, but these days a big part of movies earning included overseas earning, countries such as china, India etc have become one of the biggest earning opportunities for movie makers. As it is hard to find such information but if available can significantly impact the model.

We haven't use the actors and actress involved in the movies as it would require sophisticated models to judge the worth of an actor which changes from year to year for example, in early 2000's Adam Sandler was a pretty high worth actor by fast forward to 2018 and today adam sandler do not have the same worth as before. Therefore, even though actors involved are important part of movie revenues but I haven't used them here.

Data Pre-Processing:

Data Preprocessing in an important data mining step used to convert raw data to usable format. Often data is inconsistent with missing values, and contains many errors.

Steps Involved:

1. ***Type Adjustment***: Converting each column in proper format ie. earning in numeric, director, movie name, rating etc in character.

2. ***Handling missing values:*** All not movies have all the required data each in turn creates missing values in certain column. Currently I had multiple missing value in the dataset shown below.

```
> # finding number of NA values in each column
> colSums(is.na(Moviedf_final))
                          Rank                     Movie.Name                    Studio.Name
                             0                              0                              0
 Total.Movie.Earning.million                        Theater Earning.Opening.Day.million.
                             0                              0                              0
                          Year                          Rated                          Genre
                             0                              1                             11
                   Runtime.min                       Director                 ImdB.names.used
                            16                             27                              0
```

   As the director name is not director used in the model and data exploration therefore, all rows with 'NA' values except in the director's column are being removed.

3. ***Hot Encoding:*** In order to use different attributes such as Rating, Genre. I have created dummy variables (binary variable 0 & 1) for each of them.

Family	Fantasy	Horror	Mystery	Romance	Sci-Fi	Short	Genre
0	0	0	0	0	0	0	Action, Adventure, Sci-Fi
0	0	0	0	0	0	0	Action, Adventure, Fantasy, Sci-Fi
0	0	0	0	0	0	0	Animation, Action, Adventure, Comedy, Family, S
0	0	0	0	0	0	0	Action, Adventure, Sci-Fi
0	0	0	0	0	0	0	Action, Adventure, Comedy, Sci-Fi
0	0	0	0	0	0	0	Action, Adventure, Thriller
0	0	0	0	0	0	0	Action, Adventure, Comedy, Sci-Fi
0	0	0	0	0	0	0	Action, Adventure, Fantasy, Sci-Fi
0	0	0	0	0	0	0	Action, Sci-Fi
0	0	0	0	0	0	0	Animation, Comedy, Family, Fantasy
0	0	0	0	0	0	0	Drama, Music, Romance
0	0	0	0	0	0	0	Drama, Horror, Mystery, Sci-Fi, Thriller
0	0	0	0	0	0	0	Comedy, Romance
0	0	0	0	0	0	0	Biography, Drama, Music
0	0	0	0	0	0	0	Animation, Adventure, Comedy, Family, Fantasy
0	0	1	0	0	0	0	Horror, Thriller
0	0	0	0	0	0	0	Action, Horror, Sci-Fi, Thriller

4. ***Merging:*** I have merged the yearly data with movie dataset.

5. ***Inflation Adjustment:*** As the data is spread across 18 years so it is not accurate to compare the earning of a movie in 2002 wrt 2018 due to inflation. Therefore, I have used the average ticket price to find the inflation in movie earning (as movie ticket directly affect the earning of a movie). Using the inflation factor I find the adjusted earning for each movie.

Data Exploration and Findings:

I have used R to explore and analyze the dataset. Following are few key finding:

1. *Top Earning movies of 21st Century (i.e. 2001-2018)* ::
   "***Star Wars: The Force Awakens***" is the biggest earner of all movies released in the 21st century followed by "Avatar" (after inflation adjusting all earning data)

```
> Moviedf_final3[order(Moviedf_final3$Adj.Total.Movie.Earning.million, decreasing = TRUE),]$Movie.Name[1:10]
 [1] "Star Wars: The Force Awakens" "Avatar"                       "Marvel's The Avengers"
 [4] "Jurassic World"               "Black Panther"                "The Dark Knight"
 [7] "Avengers: Infinity War"       "Shrek 2"                      "Spider-Man"
[10] "Star Wars: The Last Jedi"
```

2. *Highest Grossing Year for movie earning in 21st Century (i.e. 2001-2018)* ::

```
> best.earning.yr[order(best.earning.yr$Adj.Total.Gross.Earning.million, decreasing = TRUE),]$Year[1:5]
[1] 2002 2003 2004 2001 2009
```

   '2002' is the highest grossing earning year with movies such as "Spider-Man", "The Lord of the Rings: The Two Towers", "Star Wars: Episode II - Attack of the Clones" & "Harry Potter and the Chamber of Secrets". It is followed by 2003 which has "The Lord of the Rings: The Return of the King" as its highest grossing movie.

3. *Highest Earning Studio* ::

```
> highest.earning.studio[order(highest.earning.studio$`Total Earning`, decreasing = TRUE),]$Studio[1:5]
[1] "BV"   "WB"   "Uni." "Fox"  "Sony"
```

   BV : Buena Vista, is a subsidiary of Disney studio, the Walt Disney Studio. (https://en.wikipedia.org/wiki/Buena_Vista_(brand)) followed by Warner Brothers, Universal, Fox and Sony.

4. *Highest Earning Genre* ::

```
> highest.earning.genre[order(highest.earning.genre$x, decreasing = TRUE),]$Genre[1:5]
[1] "Adventure" "Comedy"    "Action"    "Drama"     "Fantasy"
```

   'Adventure' genre movies proved to be highest earner movies followed by Comedy and Action.

5. *Average runtime for movies for each studio* ::

```
> studio.runtime[1:5,]
   Studio Avg.Runtime
8      BV    110.7611
56     WB    114.7973
53   Uni.    110.2661
15    Fox    107.7598
46   Sony    109.8554
```

'***110 mins***' seems to be the sweet spot for blockbuster movies for all the biggest Studio in hollywood.

6. *Highest Earning Rating for the movies* ::

```
> highest.earning.rating[order(highest.earning.rating$Earning, decreasing = TRUE),]
     Rating   Earning
1         G 115873880
4        PG 107975402
5     PG-13 100352621
2       N/A  71852142
6         R  65469002
3 NOT RATED  62245308
7     TV-14  47617067
8   UNRATED  28941795
```

'***G***' rating i.e. General Admission is the common rating provided to highest grossing movies. Followed by '***PG***' (i.e. Parental Guidance) rating.

7. *Average runtime for movies in different ratings* ::

```
> merge(rating.info, rating.movies, by='Rating')
     Rating   Earning Avg. Runtime Number of Movies
1         G 115873880     93.55556               45
2       N/A  71852142     72.39474               38
3 NOT RATED  62245308     44.50000                2
4        PG 107975402    100.74766              321
5     PG-13 100352621    113.41596              827
6         R  65469002    112.93855              537
7     TV-14  47617067     60.00000                1
8   UNRATED  28941795     40.00000                2
```

Most of the movies consist of PG, PG-13 and R rating with average run time around 2 hours even though 'G' has the highest earning.

8. *Biggest Opening day earner of the 21st Century* ::

```
> Moviedf_final3[order(Moviedf_final3$Adj.Earning.Opening.Day.million, decreasing = TRUE),]$Movie.Name[1:5]
[1] "Star Wars: The Force Awakens" "Avatar"                       "Marvel's The Avengers"
[4] "Jurassic World"               "Black Panther"
```

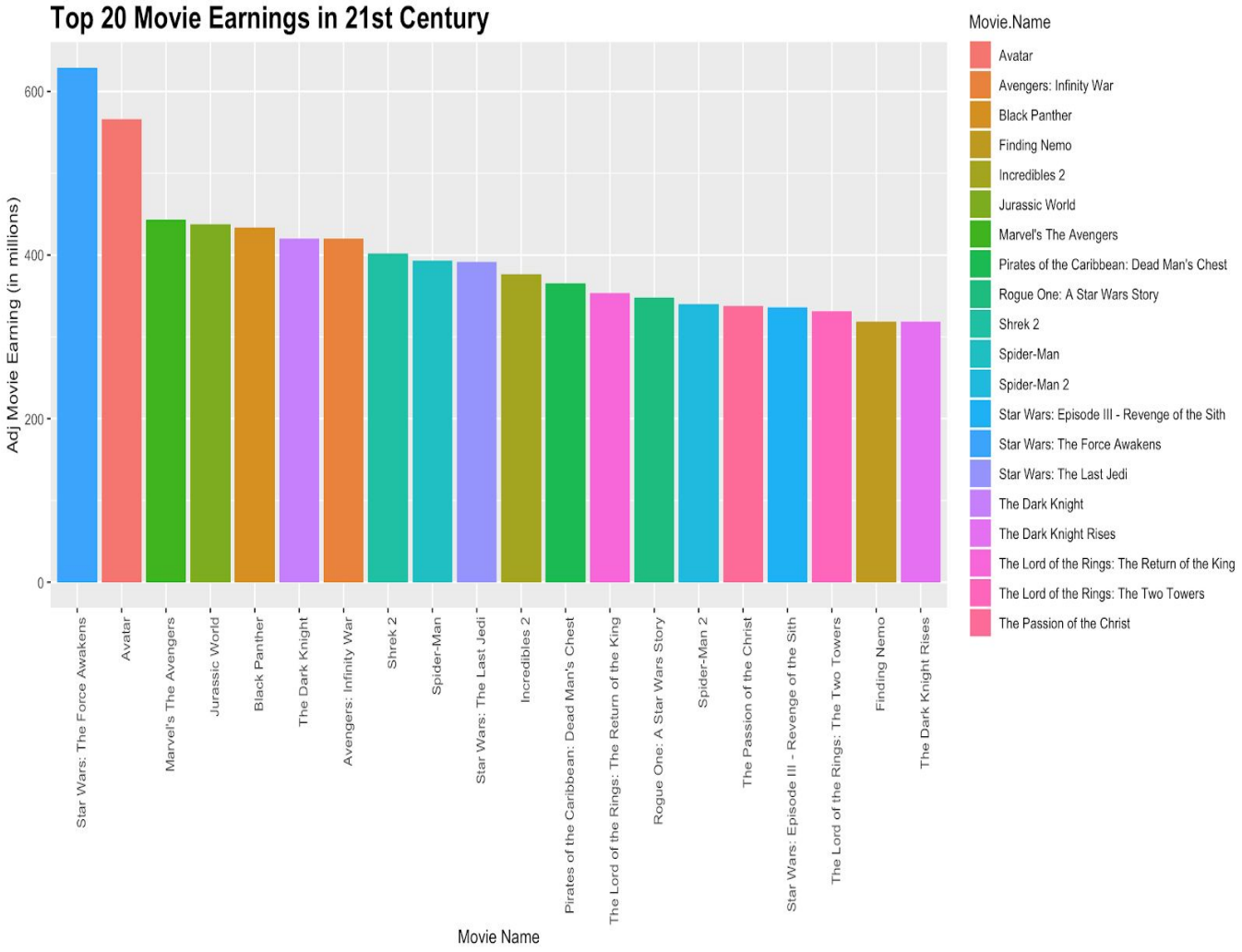
The highest earner are also the ones with the biggest opening day which suggests correlation between the two.

9. Mean Earning from top 100 movies each year ::

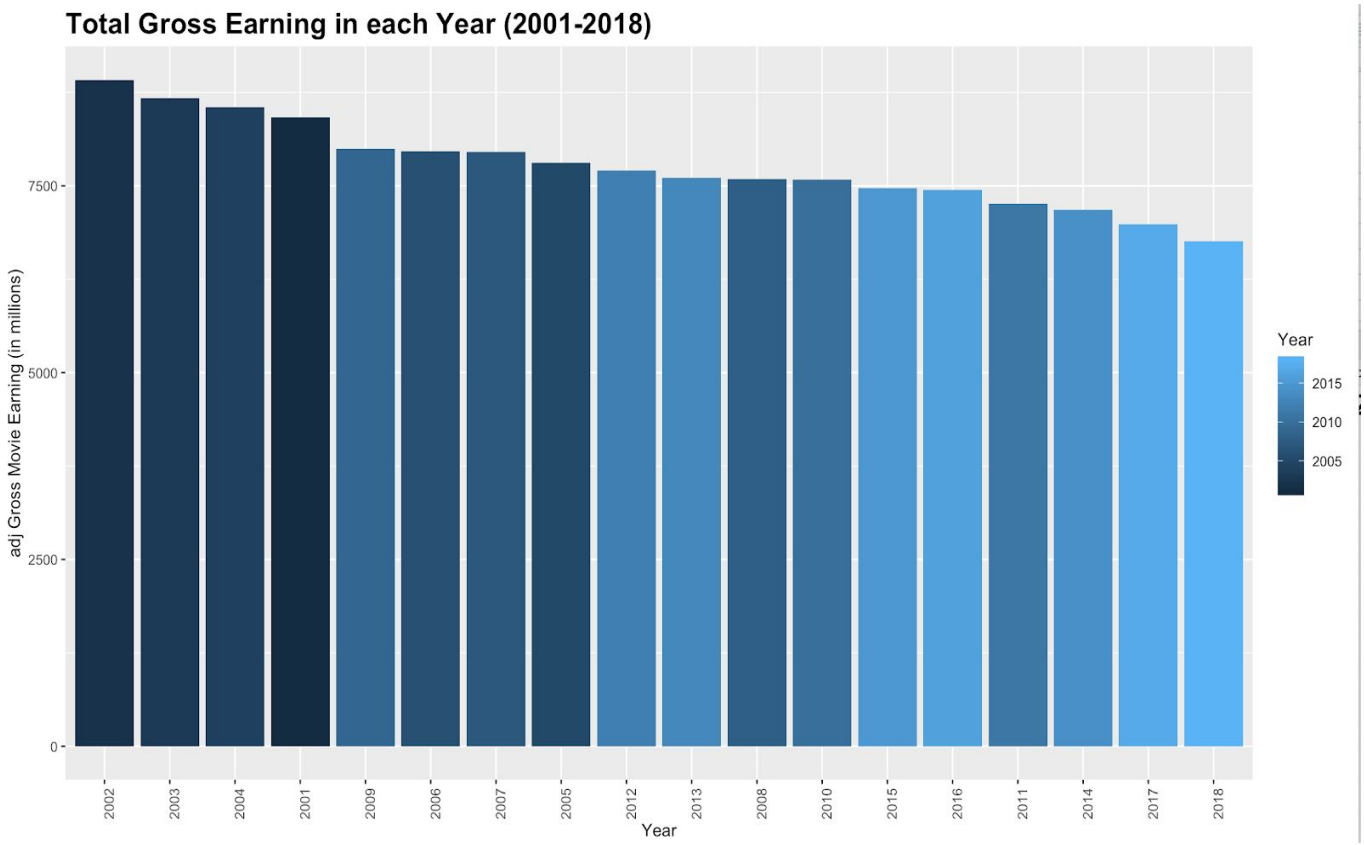
```
> mean(Moviedf_final3$Adj.Total.Movie.Earning.million)
[1] 69829871
```

Visualizations :

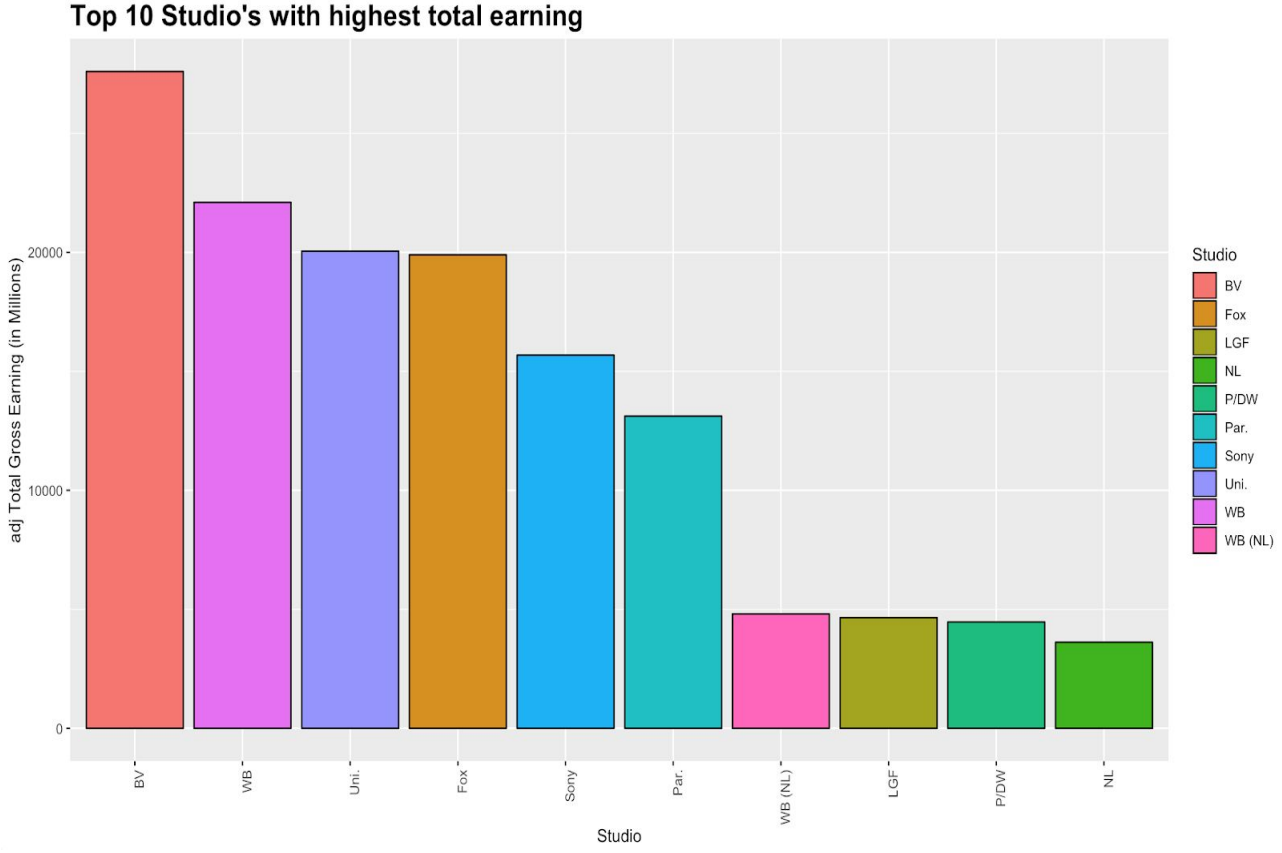
1. Top 20 movies earning in 21 century:



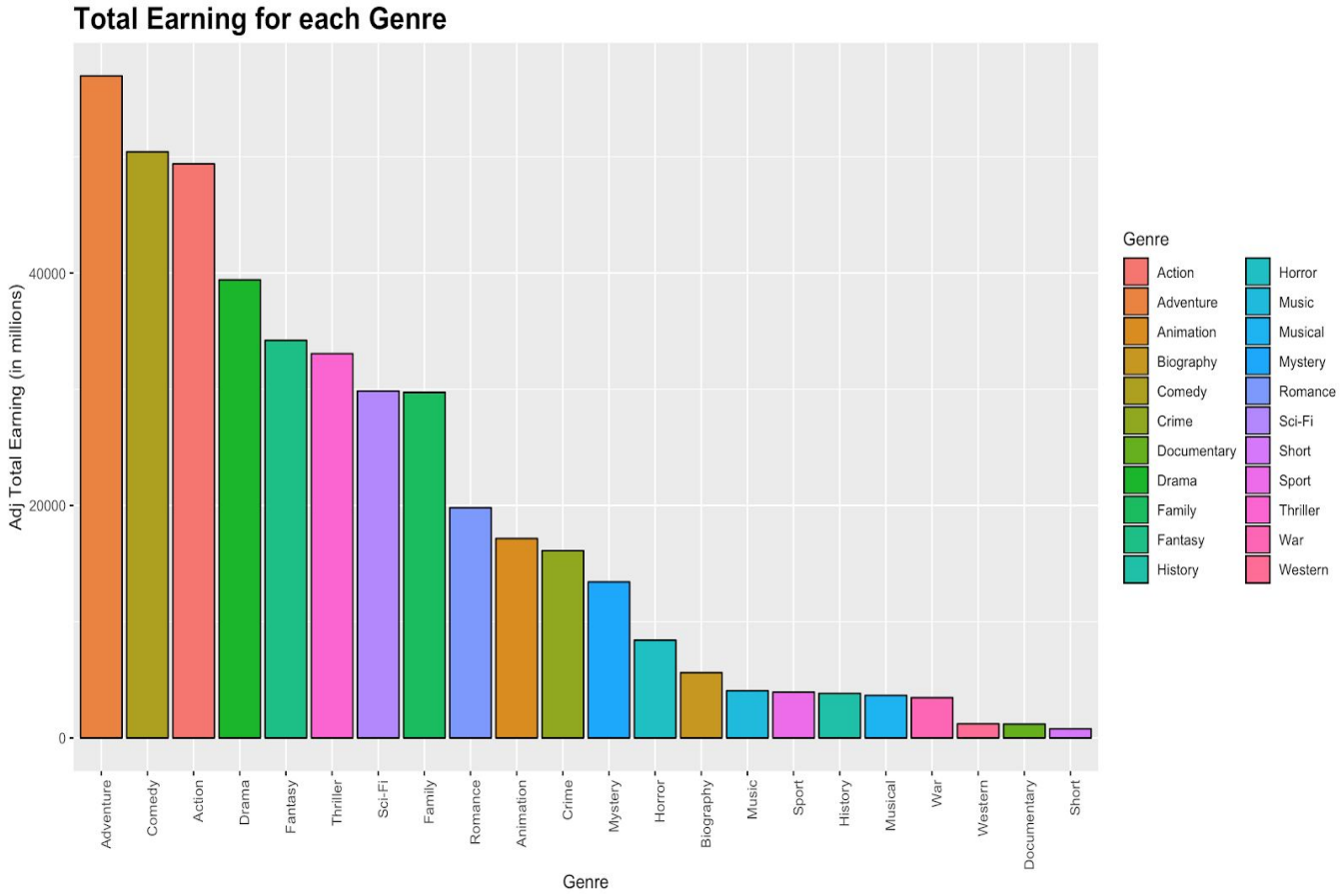
2. Total gross earning in each year:



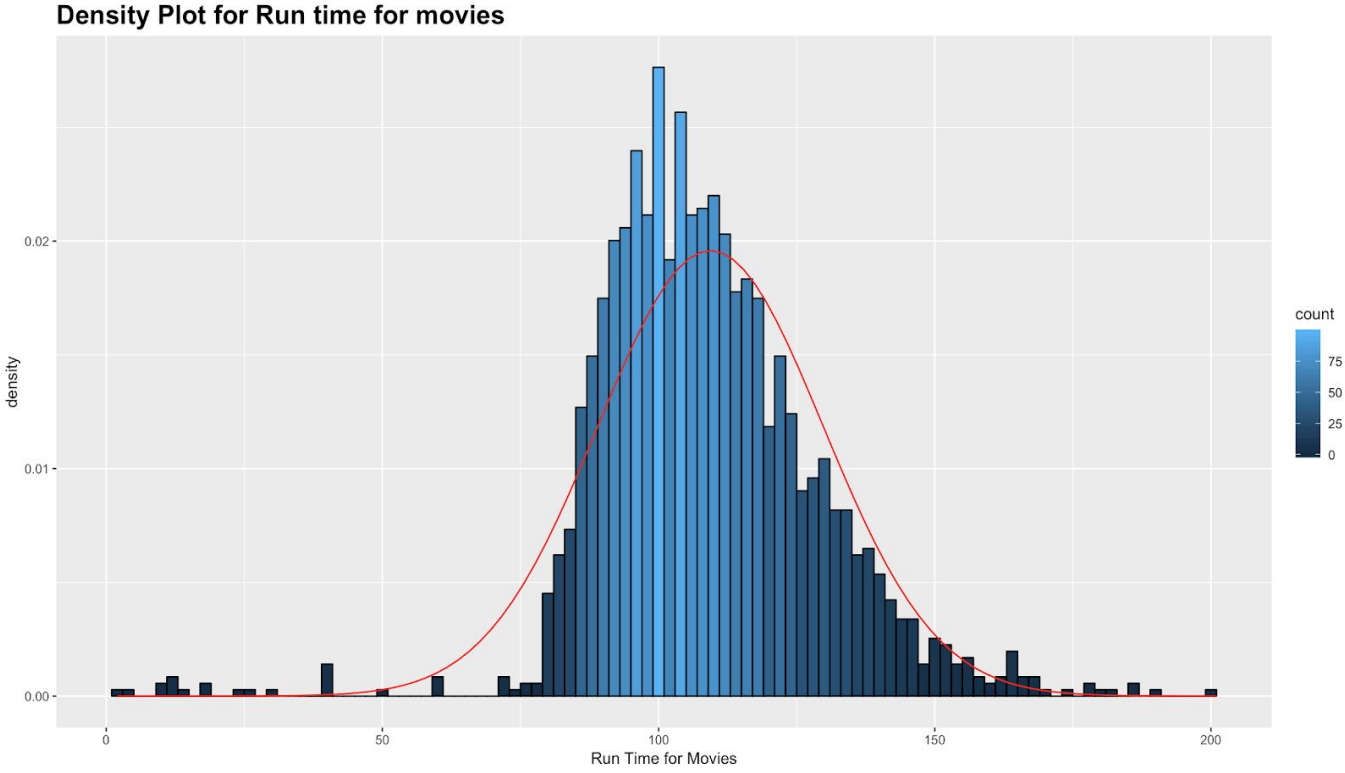
3. Top 10 studios with highest earning:



4. Total earning for each genre:



5. Density Plot for Run time for Movies



By looking at the density plot we can say that 110 is the mean runtime for movies.

Conclusion/Interpretation of results:

The most challenging part of this project is to scrap the data from website and transform it into usable format. The results obtained above are the general analysis which helps me to find the top earning movie, which genre earns the highest for the particular year, which year is the highest earning year for the movies companies etc. Many more questions/observations can also be answered using this dataset.

Future work:

1. I can collect more data from different website also like Netflix.com, popcornflix.com etc which can add more details about movies. 2.
2. I can add actors, actresses with directors too because movie box office can also be influenced by the combination of actors and directors.
3. Rather than top 100 movies, analysis could be extended for all movies.
4. Budget of the movie is also an influential factor.
5. Season can also be considered as important factors as most of the successful movies are released during the summer or winter break.
6. Critic opinion can also be considered as it is also one of the factor of earning. Tomatometer and Tomato User meter are critics and user rating and it was surprising to observe that both had a similar effect, which mean audience still agree with the critics opinions.

References:

1. https://www.imdb.com/
2. https://www.rottentomatoes.com/
3. https://www.boxofficemojo.com/
4. http://www.natoonline.org/
5. http://www.omdbapi.com/
6.https://stackoverflow.com/questions/52631921/find-unique-values-in-a-character-vector-separated-by-commas-and-then-one-hot-en
7.https://stackoverflow.com/questions/4862178/remove-rows-with-all-or-some-nas-missing-values-in-data-frame
8.https://stackoverflow.com/questions/26273663/r-how-to-total-the-number-of-na-in-each-col-of-data-frame